Day 1: Overview and Introduction to Data Science

ME414: Introduction to Data Science and Big Data Analytics
LSE Methods Summer Programme
14 August 2017

Course outline and logistics

Pitch of this course

- Whom this class is for
    - (typically) graduate-level applied researchers with some prior experience in quantitative methods
    - those requiring the fundamentals of data science
    - people working with typically large datasets and databases
    - those seeking a practical introduction to data analysis using R
- Prerequisites
    - at least one prior course in quantitative methods or statistics
    - familiarity, or willingness to learn, R
    - ability to use a text editor and spreadsheet is helpful
    - (optional) ability to process data files in a programming language such as Python

Learning objectives

- an overview of data science and the challenges of working with big data
- learning the R statistical package and practical methods for manipulating and analyzing data
- how to acquire, process, store, and use data, both structured and unstructured
- supervised learning approaches, including linear and logistic regression, and decision trees,
- unsupervised learning approaches, including clustering, and principal components analysis
- quantitative methods of text analysis, including mining social media and other online resources

Who we are

Ken Benoit, London School of Economics, kbenoit@lse.ac.uk

Jack Blumenau, University College London, j.blumenau@lse.ac.uk

Slava Mikhaylov, University of Essex, s.mikhaylov@ucl.ac.uk

Altaf Ali, University College London, altaf.ali@ucl.ac.uk

Class schedule

	Lecture		Computer Class 32L LG.05B				Evening
Monday 14 August	CLM 5.02 09:30-12:00	Lunch break	Class Group 1 13:00 - 14:30	Class Group 2 14:30-16:00	Class Group 3 16:00-17:30		Welcome Reception from 17:30
Tuesday 15 August	CLM 5.02 09:00-12:00		Class Group 3 13:00 - 14:30	Class Group 1 14:30-16:00	Class Group 2 16:00-17:30		
Wednesday 16 August			Class Group 2 13:00 - 14:30	Class Group 3 14:30-16:00	Class Group 1 16:00-17:30		
Thursday 17 August			Class Group 1 13:00 - 14:30	Class Group 2 14:30-16:00	Class Group 3 16:00-17:30		
Friday 18 August	Wolfson Theatre 09:00-12:00		Class Group 3 13:00 - 14:30	Class Group 1 14:30-16:00	Class Group 2 16:00-17:30		Evening Reception from 17:30

	Lecture		Computer Class 32L LG.05B				Evening
Monday 21 August	Wolfson Theatre 09:00-12:00	Lunch break	Class Group 2 13:00 - 14:30	Class Group 3 14:30-16:00	Class Group 1 16:00-17:30		
Tuesday 22 August			Class Group 1 13:00 - 14:30	Class Group 2 14:30-16:00	Class Group 3 16:00-17:30		
Wednesday 23 August			Class Group 3 13:00 - 14:30	Class Group 1 14:30-16:00	Class Group 2 16:00-17:30		
Thursday 24 August			Class Group 2 13:00 - 14:30	Class Group 3 14:30-16:00	Class Group 1 16:00-17:30		
Friday 25 August			Quiet room (location to be confirmed) 14:00-16:00				Evening Reception from 16:30

CLM 5.02 is located on the fifth floor of Clement House

The Wolfson Theatre is located on the lower ground floor of the New Academic Building

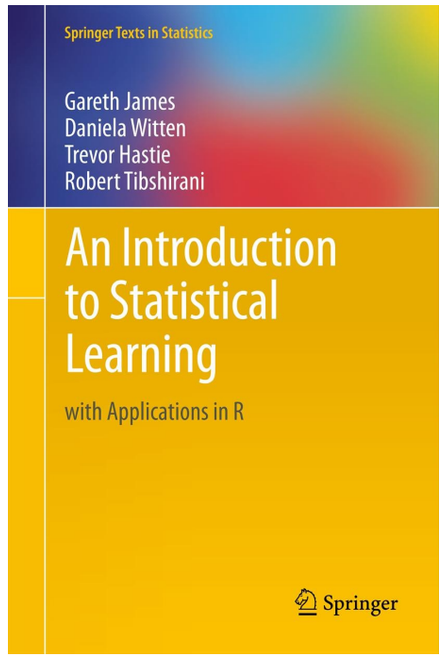
32L LG.05B is located on the lower ground floor of 32 Lincoln's Inn Fields

Essential course resources

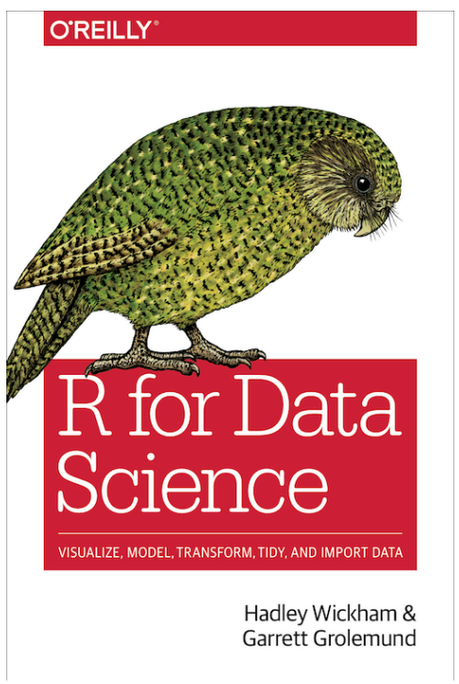
- GitHub repository: http://github.com/kbenoit/ME414
- Moodle page: https://shortcourses.lse.ac.uk/course/view.php?id=158
- Using GitHub and forking the course repo
- RStudio and GitHub integration
- How to complete (and submit) assignments. We do not mark the assignments but we walk through the solutions at the start of the lab the following day.

Course Texts

- The course will cover most of the material in ISLR. Each chapter ends with an R lab, in which examples are developed. An electronic version of this book is available for free from the authors' websites.



- For statistical learning component of the course we closely follow ISLR, including figures and lecture materials, as made available by the authors.



- Grolemund & Wickham is excellent for anything relating to R.

Concept of Data Science

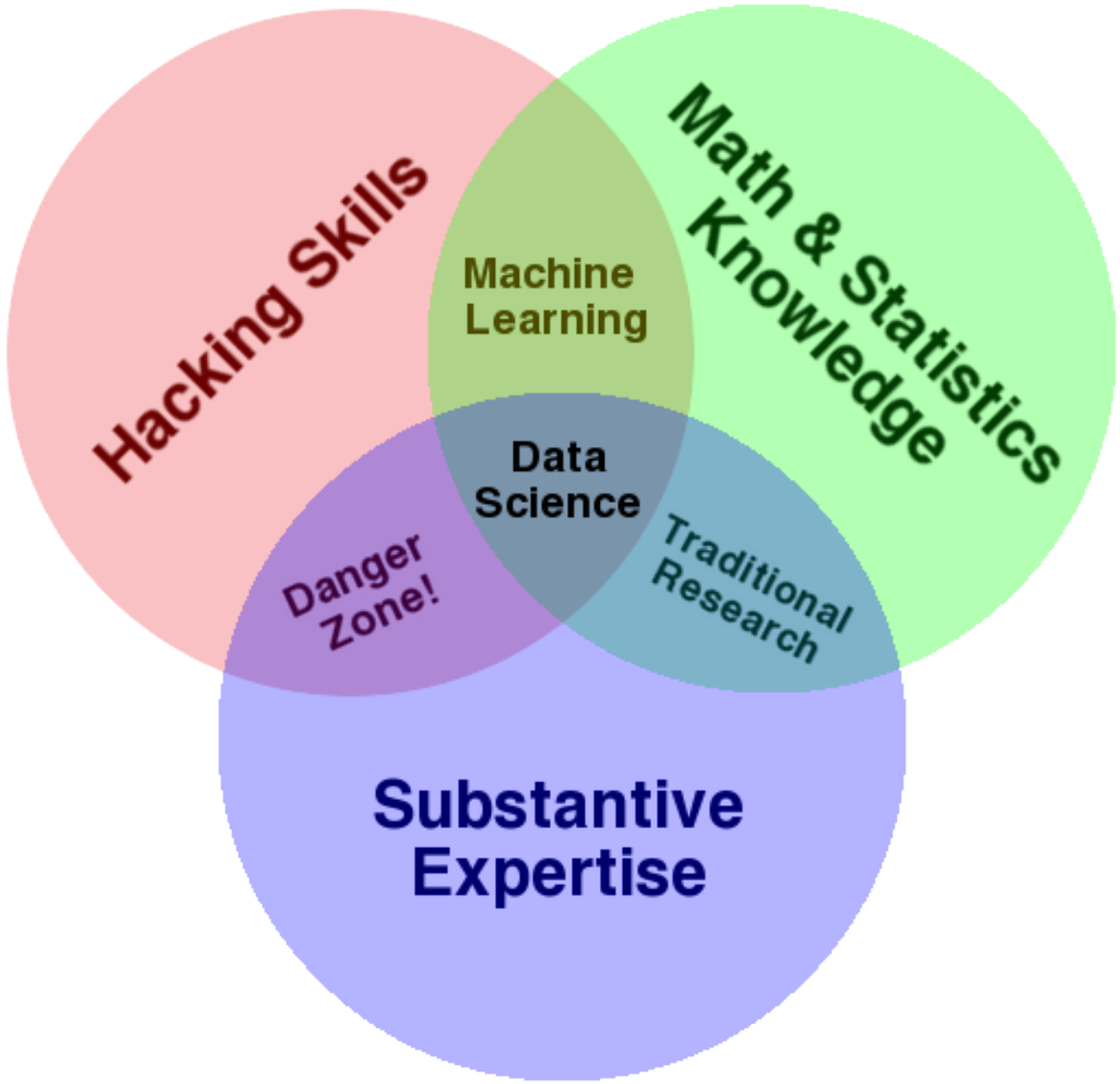
Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

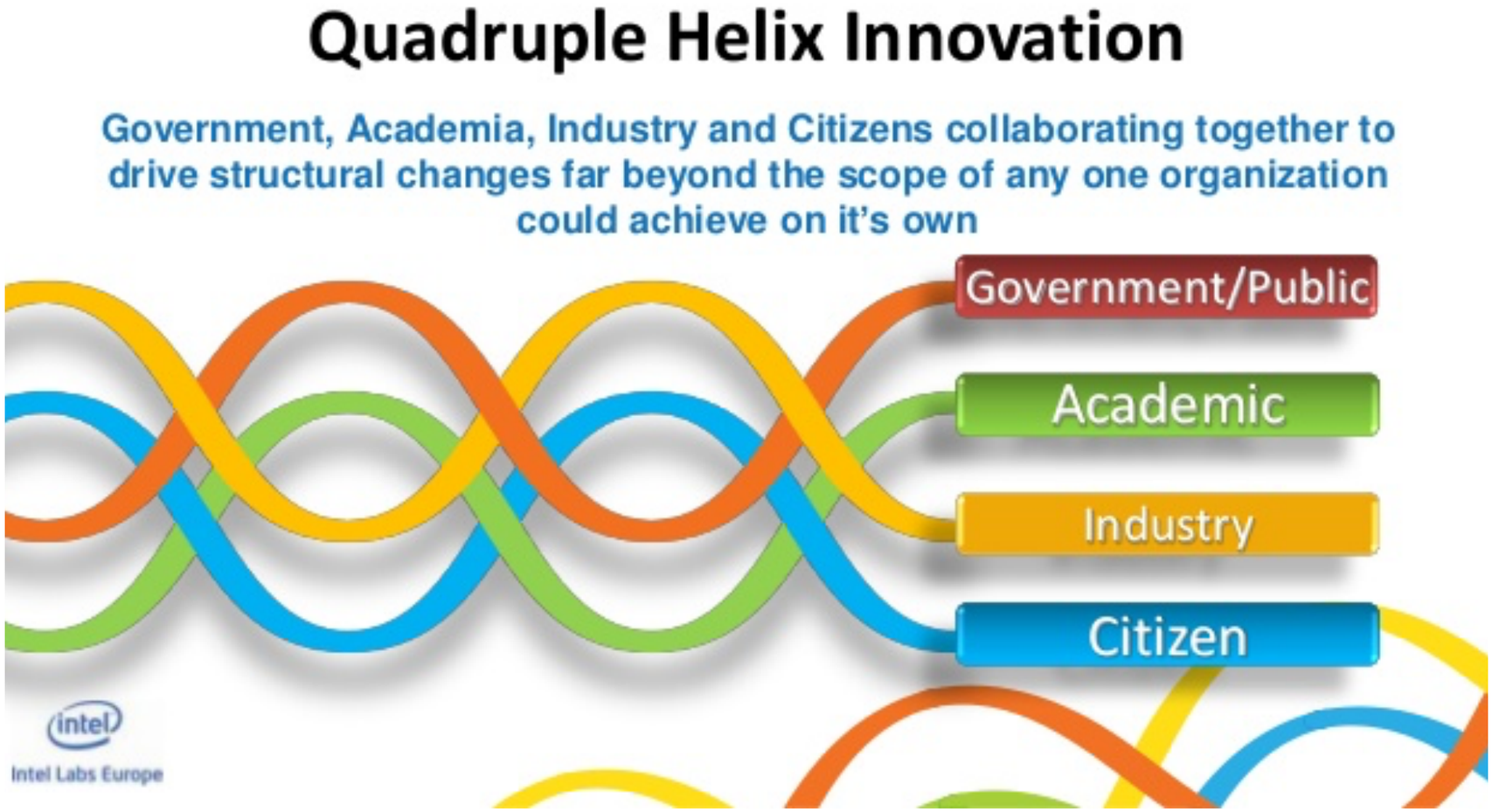
``I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?'' Hal Varian (Chief Economist at Google, 2009).

What is Data Science?



Drew Conway

System approach to data science



"Research in Big Data should be grounded in the quadruple helix model where civil society joins with business, academia, and government sectors to drive changes far beyond the scope of what any organization can do on their own."

Intel Corp policy position paper on Big Data

Practical perspective

- Data Science is the art of turning data into actions

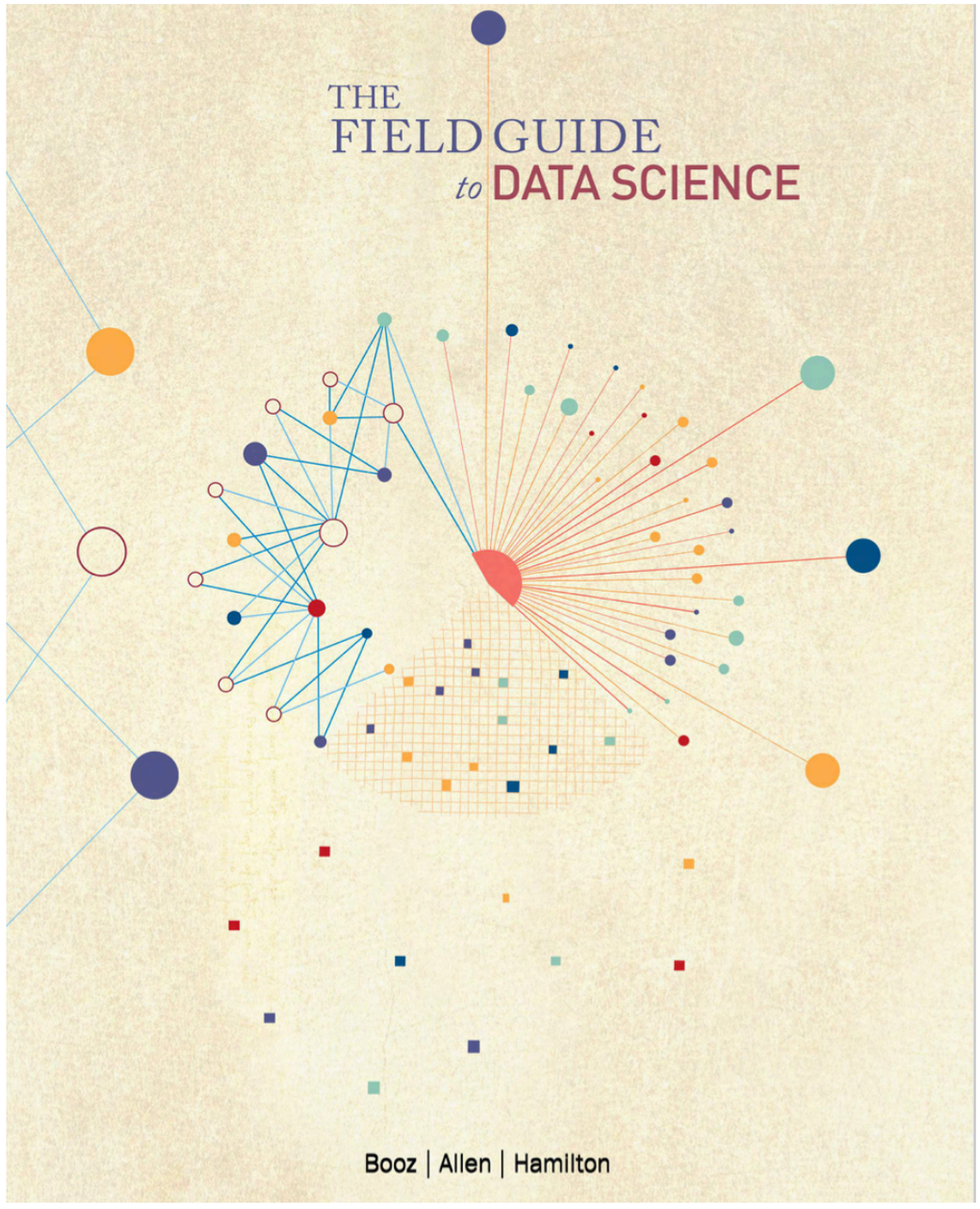
- It's all about the tradecraft.

- Tradecraft is the process, tools and technologies for humans and computers to work together to transform data into insights.

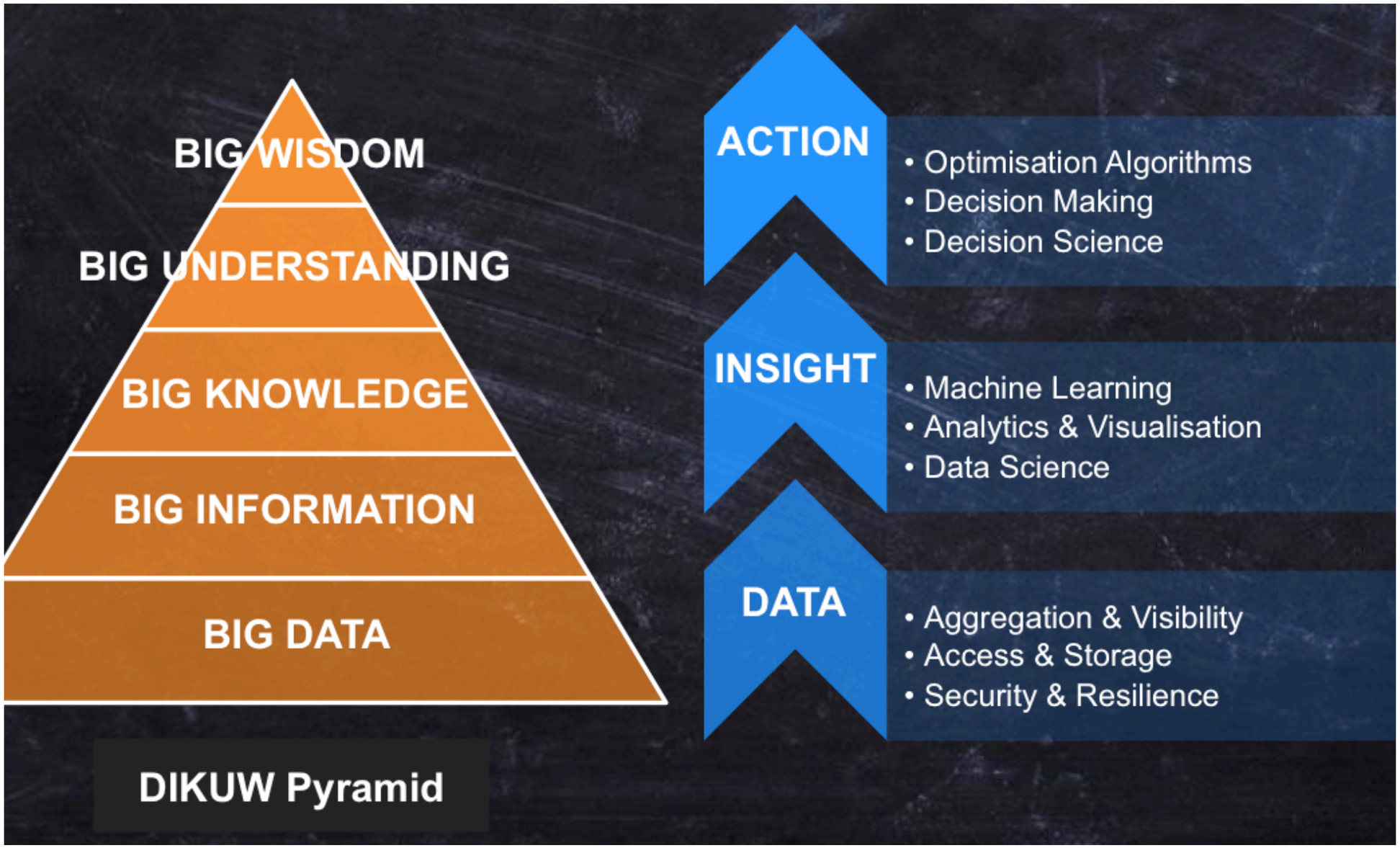
Practical perspective

- Data Science tradecraft creates data products

- Data products provide actionable information without exposing decision makers to the underlying data or analytics.



DIKUW Pyramid



Inductive and deductive reasoning

- Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning

- This is a fundamental change from traditional analysis approaches.

- Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths.

- Models of reality no longer need to be static.

- They are constantly tested, updated and improved until better models are found.

LOOKING BACKWARD AND FORWARD

FIRST THERE WAS BUSINESS INTELLIGENCE	NOW WE'VE ADDED DATA SCIENCE
Deductive Reasoning	Inductive and Deductive Reasoning
Backward Looking	Forward Looking
Slice and Dice Data	Interact with Data
Warehoused and Siloed Data	Distributed, Real Time Data
Analyze the Past, Guess the Future	Predict and Advise
Creates Reports	Creates Data Products
Analytic Output	Answer Questions and Create New Ones
	Actionable Answer

Source: Booz Allen Hamilton

Practical perspective

- Data Science is necessary for companies to stay with the pack and compete in the future.

- Data Science capabilities can be built over time.

- Data Science is a different kind of team sport.

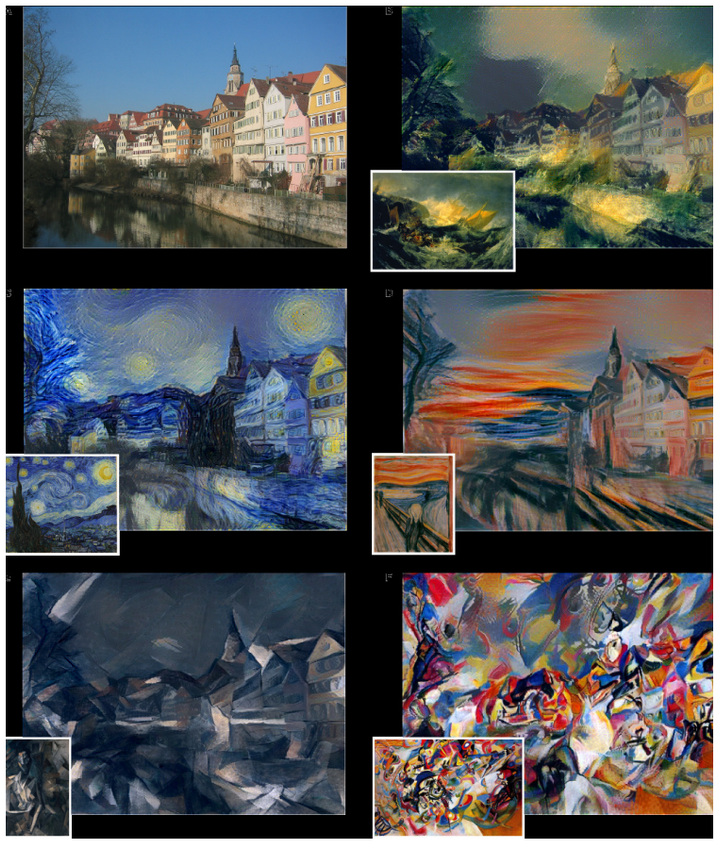
Principles of Data Science

Data Science principles

- Be willing to fail.
- Fail often and learn quickly.
- Keep the goal in mind.
- Dedication and focus lead to success.

Learning quickly

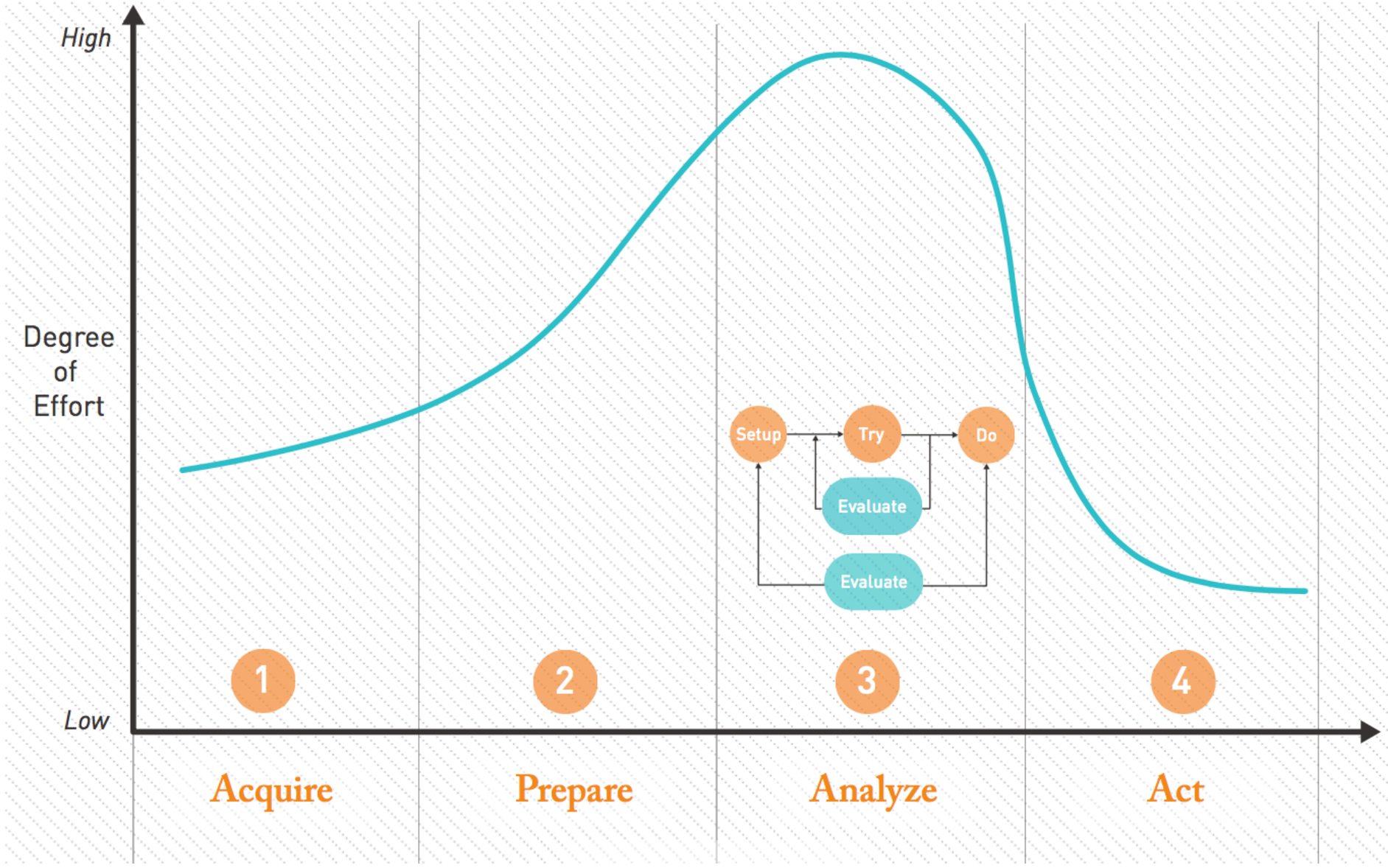
- Leon A. Gatys, Alexander S. Ecker, Matthias Bethge. ``A Neural Algorithm of Artistic Style.'' arXiv:1508.06576. September 2015.



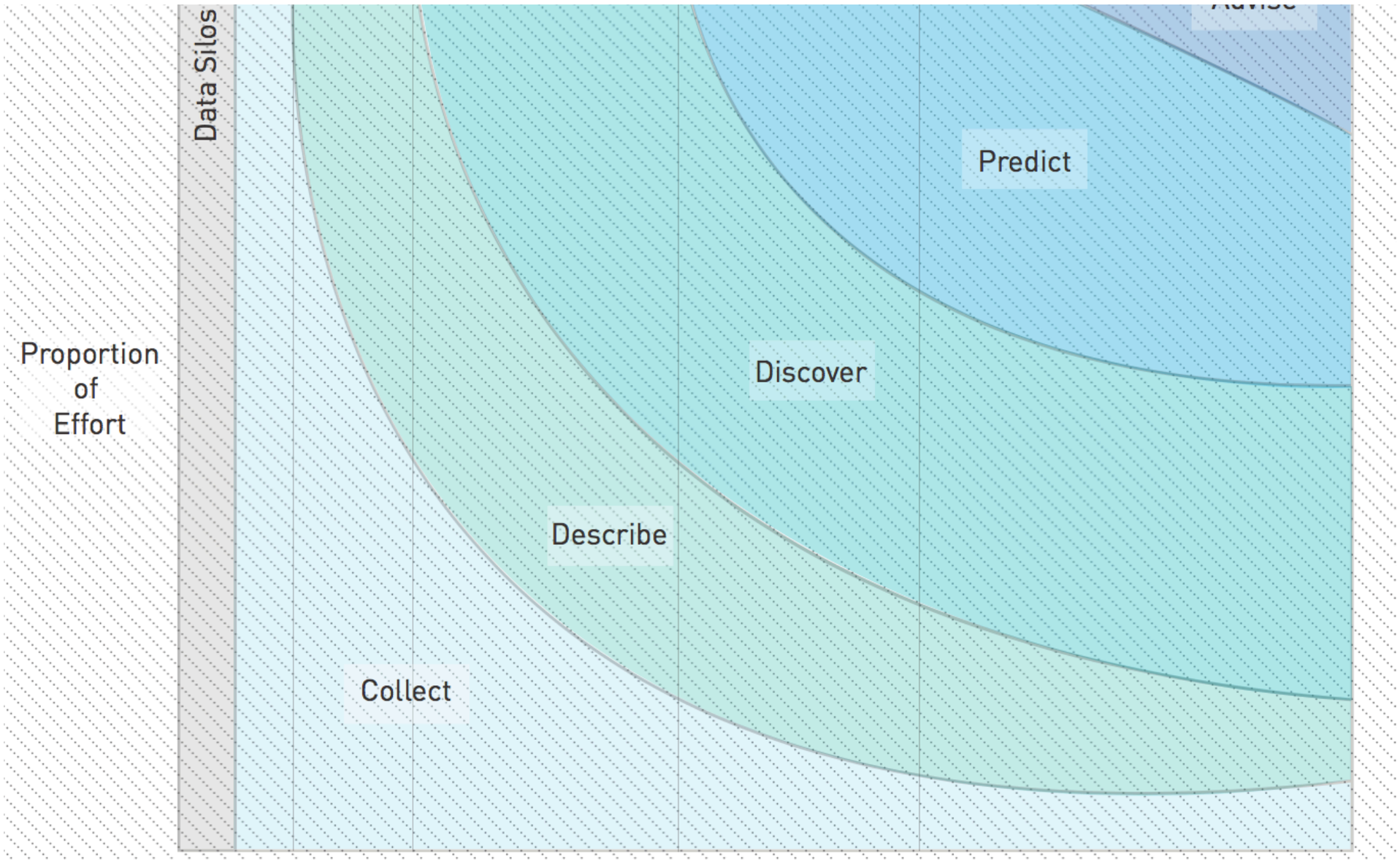
- Prisma and Convolutional Neural Networks: June 2016.

Practice of Data Science

Data science workflow

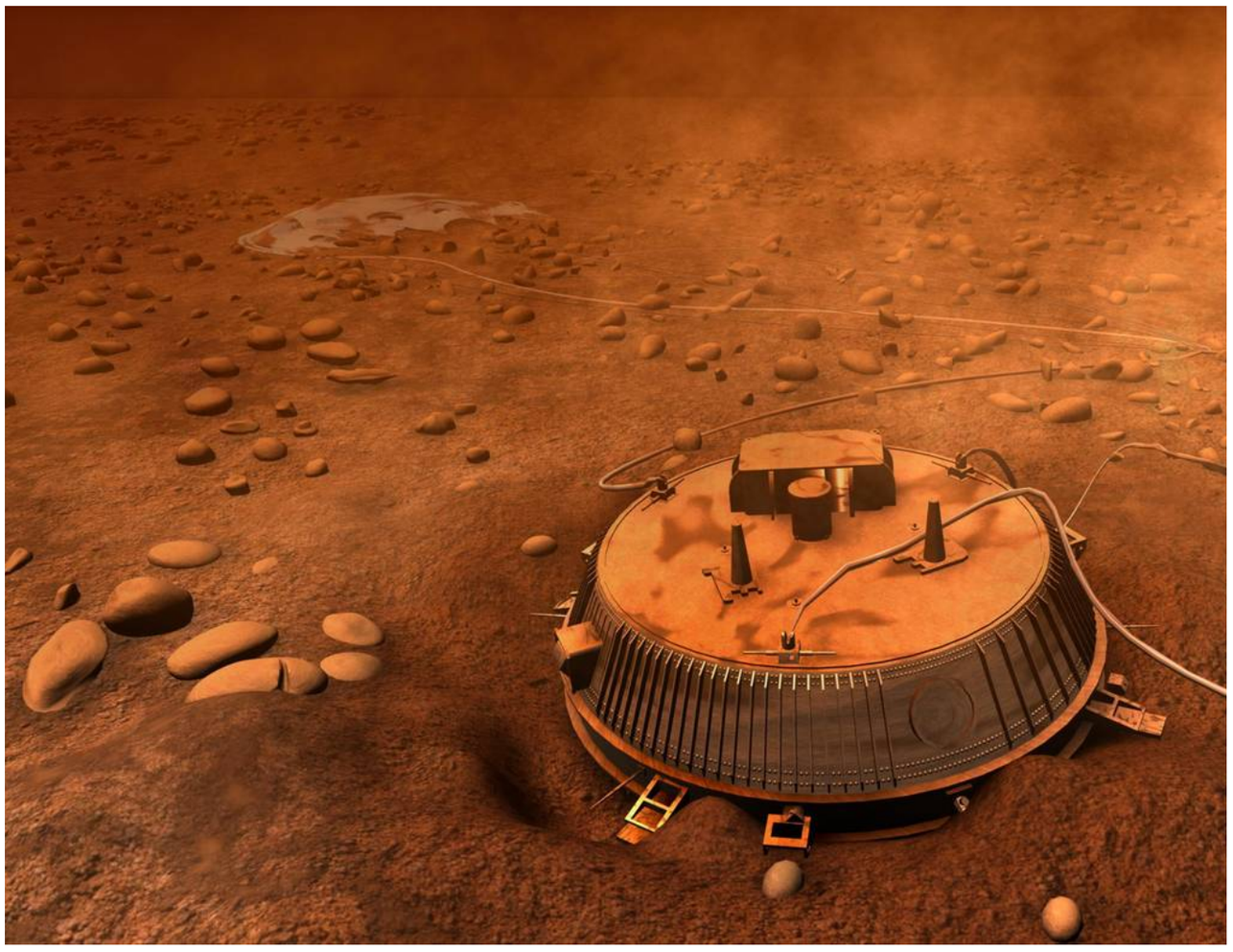


Data science in organisations



Unlocking value from data: Saturn and Cassini-Huygens mission

- "Bouncing on Titan: Motion of the Huygens Probe in the Seconds After Landing" by Stefan Schroeder et al (2 Feb 2017, ArXiv)

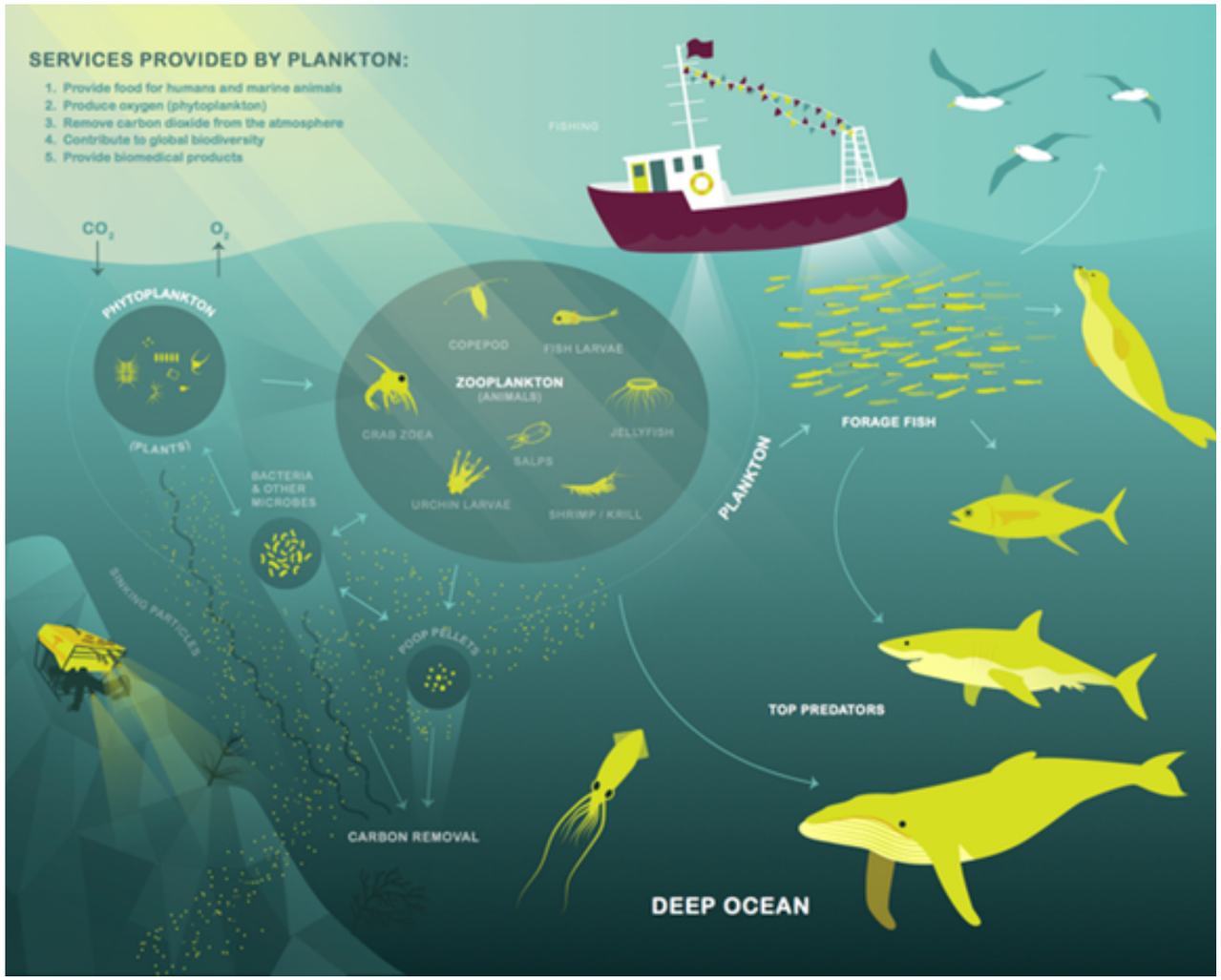


- Re-analyzing data from 2005. But also including (linking) data from a variety of other instruments. Get insights into surface structure.

- The way it bounced points to damp sand. Methane lakes.

Kaggle Competitions

- National Data Science Bowl - Predict ocean health, one plankton at a time.
- Passenger Screening Algorithm Challenge - Improve the accuracy of the US Department of Homeland Security's threat recognition algorithms
- Chicago Department of Public Health West Nile Virus Prediction - Predict West Nile virus in mosquitos across the city of Chicago.
- Redefining Cancer Treatment - Predict the effect of Genetic Variants to enable Personalized Medicine
- Genentech Flu Forecasting - Predict when, where and how strong the flue will be.
- Mercedes-Benz Greener Manufacturing - Can you cut the time a Mercedes-Benz spends on the test bench?



Governance

Data Science Ethical Framework

Cabinet Office

19 May 2016

Principles

1. Start with clear user need and public benefit

2. Use data and tools which have the minimum intrusion necessary

3. Create robust data science models

4. Be alert to public perceptions

5. Be as open and accountable as possible

6. Keep data secure

1 Start with clear user need and public benefit

- Data science offers huge opportunities to create evidence for policymaking, and make quicker and more accurate operational decisions.

- Being clear about the public benefit will help you justify the sensitivity of the data (principle 2) and the method that you want to use (principle 3).

2 Use data and tools which have the minimum intrusion necessary

- You should always use the minimum data necessary to achieve the public benefit.

- Sometimes you will need to use sensitive personal data.

- There are steps that you can take to safeguard people's privacy e.g. de-identifying or aggregating data to higher levels, querying against datasets or using synthetic data.

3 Create robust data science models

- Good machine learning models can analyse far larger amounts of data far more quickly and accurately than traditional methods.

- Think through the quality and representativeness of the data, flag if algorithms are using protected characteristics (e.g. ethnicity) to make decisions, and think through unintended consequences.

- Complex decisions may well need the wider knowledge of policy or operational experts.

4 Be alert to public perceptions

- The Data Protection Act requires you to have an understanding of how people would reasonably expect their personal data to be used.

- You need to be aware of shifting public perceptions.

- Social media data, commercial data and data scraped from the web allow us to understand more about the world, but come with different terms and conditions and levels of consent.

5 Be as open and accountable as possible

- Being open allows us to talk about the public benefit of data science.

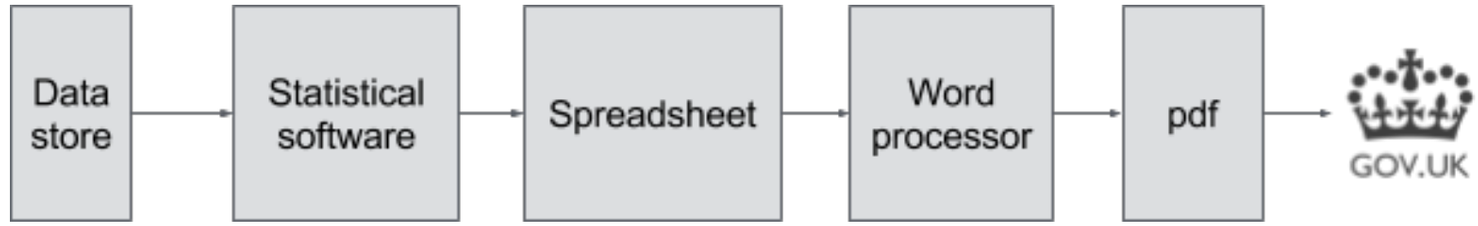
- Be as open as you can about the tools, data and algorithms (unless doing so would jeopardise the aim, e.g. fraud).

- Provide explanations in plain English and give people recourse to decisions which they think are incorrectly made.

- Make sure your project has oversight and accountability built in throughout.

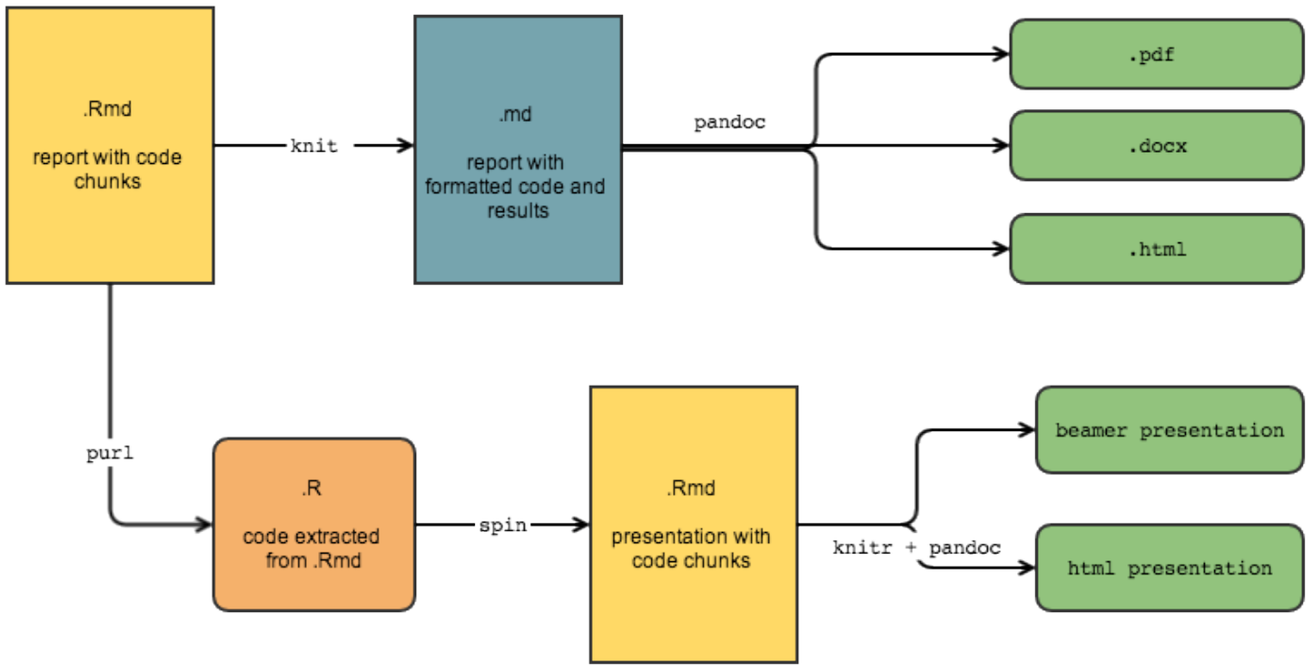
Reproducible Analytical Pipelines

- Open source rather than proprietary (R and Python)
- Version control, packaging code
- Procedural programming and unit testing
- Dependency management
- Data testing



Reproducible Analytical Pipelines and RMarkdown Reporting

- GDS RMarkdown reporting
- UK GDS on GitHub



6 Keep data secure

- We know that the public are justifiably concerned about their data being lost or stolen.

- Government has a statutory duty to protect the public's data and as such it is vital that appropriate security measures are in place.

Quick checklist

Tick where you are on the scale

1. Start with clear user need and public benefit

A. How does the department and public benefit?

- High public benefit (to society or to an individual)
- Medium public benefit (to society or to an individual)
- Low public benefit (to society or to an individual)

2. Use data and tools which have the minimum intrusion necessary

B. How intrusive and identifiable is the data you are working with?

- Non-personal and therefore non-identifiable
- Personal but non-sensitive
- Personal, sensitive data which could be inferred or directly re-identified

C. If identifying individuals, how widely are you searching personal data?

- Querying against known individuals
- Querying against a targeted group
- Speculatively searching for needle in haystack

3. Create robust data science models

D. What is the quality of the data?

- Representative and unbiased
- Historical data which is biased and excludes certain groups
- Inaccurate or missing data

E. How automated are the decisions?

- Human making decision based on analysis
- Limited human oversight but regularly checked
- No human oversight or method of checking

F. What is the risk that someone will suffer a negative unintended consequence as a result of the project?

- Low
- Medium
- High

4. Be alert to public perceptions

G. If personal data for operational purposes, how compatible was it with the reason collected?

- Very compatible
- Less compatible but fair
- Not compatible

H. Do the public agree with what you are doing?

- Yes
- Some would, some wouldn't or not sure what people think
- No, or lots would have real concerns

5. Be as open and accountable as possible

I. How open can you be about the project?

- Very open, and make open the tools and data for re-use
- Open about project but not about data/tools
- Cannot talk about project aim

J. How much oversight and accountability is there throughout the project?

- Throughout - including the decision made as a result of insight
- Only at the beginning
- None

6. Keep data secure

K. How secure is your data?

- Very secure, with restricted access to a few named individuals
- Secure and password protected
- Openly available within the department

*Not all may apply to your project

All fine? Go forward!

Some issues? Think carefully

Tricky issues? Extreme care & oversight

Tricky issues?
Extreme care & oversight

Answering these questions will also act as your Privacy Impact Assessment

1. Start with clear user need and public benefit

How does the public benefit outweigh the risks to privacy and the risk that someone will suffer an unintended negative consequence? (PIA Step 1)

Brief description of the project, including data to be used, how it will be collected and deleted. (PIA Step 2)

2. Use data and tools which have the minimal intrusion necessary

What steps are you taking to maximise the benefit of the project outcome?

What steps are you taking to minimise risks to privacy? (for example using less intrusive data, aggregating data etc)?

3. Create robust data science models

What steps have you taken to make sure the insight is as accurate as possible and there are minimal unintended consequences? (for example thinking through quality of the data, human oversight, giving people recourse)

4. Be alert to public perceptions

How have you assessed what the public or stakeholders would think of the acceptability of the project? What have you done in addition to the above to address any concerns?

Risks (PIA Step 3) and mitigating steps (PIA Step 4)

5. Be as open and accountable as possible

How are you telling people about the project and how you are managing the risks?

Who has signed this off within your organisation? Who will make sure the steps are taken and how? PIA Step 5

6. Keep data secure

What steps are you taking to keep the data secure?

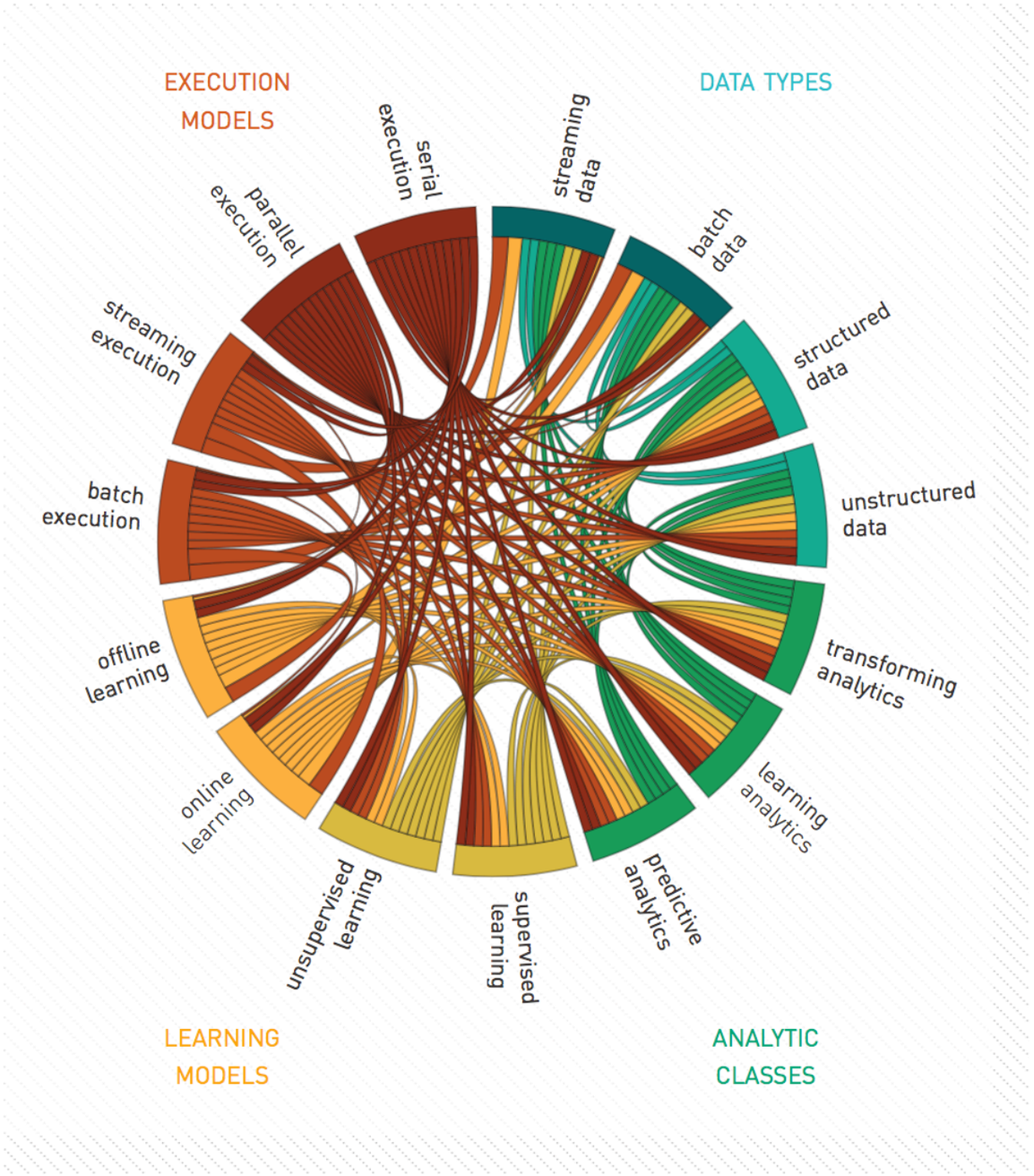
Case study

- Discussion of the case in the media
- ICO ruling
- Checking against Data Science Ethical Framework

Perceptions

- Does your assessment change if you take into account previous “issues”?
- care.data case

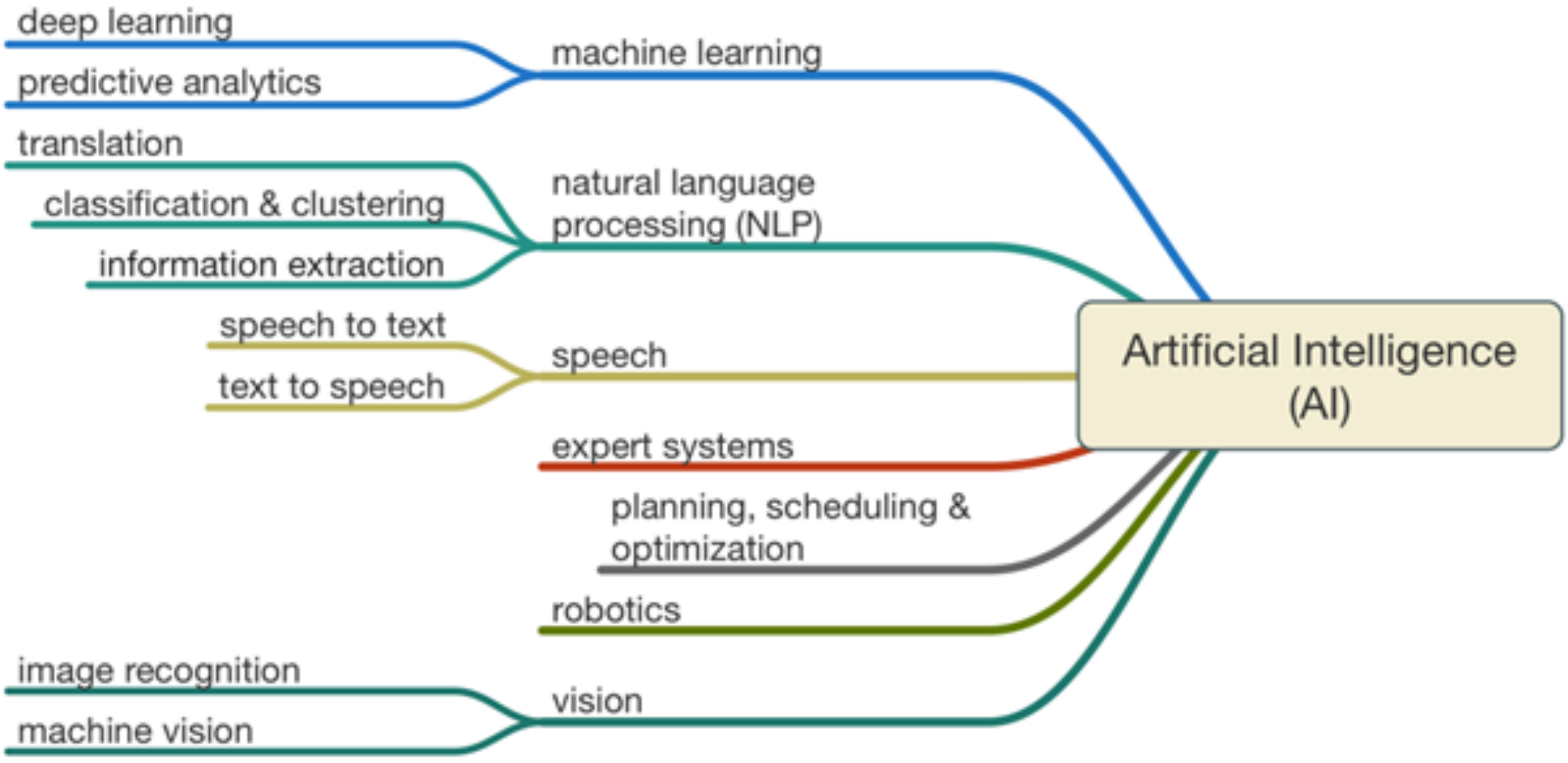
Analytics



Source: Booz Allen Hamilton

Interconnection Among the Component Parts of Data Science

Analytics components



Neota Logic

Main models and approaches